MS-E2112 Multivariate Statistical Analysis – 2016
Exam
Answer to all the questions.

You are allowed to have pens and pencils, an eraser and a ruler, a basic calculator (not graphical) and one A4 note (handwritten, text on one side only, name on the top right corner).

1. True or False (6 p.)

   Determine whether the statement is true or false. A statement is true if it is always true — otherwise it is false. (Every correct answer +3/8 p., every wrong answer -3/8 p., no answer 0 p.)

   (a) PCA transformation is invariant under affine transformations.

   (b) PCA is sensitive to heterogenous scaling of the variables.

   (c) If the influence function of a functional Q is bounded (with respect to L_2 norm), then the asymptotical breakdown point of Q can not be 0.

   (d) The asymptotical breakdown point of the mean vector is 1/2.

   (e) The componentwise multivariate median is affine equivariant.

   (f) Under the assumption of multivariate ellipticity, all affine equivariant scatter functionals are proportional.

   (g) In bivariate correspondence analysis, PCA is applied to scaled and shifted contingency tables of relative frequencies.

   (h) Multiple correspondence analysis (MCA) is based on applying bivariate correspondence analysis on the so called complete disjunctive table.

   (i) Whereas PCA relies on euclidian distances, MCA relies on chi-square distances.

   (j) In MCA, rare modalities have neglible/small effect on the analysis.

   (k) Canonical correlation analysis focuses on relationships within groups of variables.

   (l) Canonical correlation analysis is symmetric on the two groups of variables.

(m) Fisher's linear discriminant analysis is based on maximizing the ratio of dispersions between groups and within group dispersions.

(n) Discriminant analysis is a method for splitting a set of individuals into unknown homogenous groups.

(o) The initial K centers do not have an effect on the results of the Moving centers clustering methods (K-means clustering methods).

(p) If the sample size n is large, clustering is usually performed by considering all the possible partitions of the n data points into K classes, $K = 1, 2, ..., n$.

Statement	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p
True																
False																

2. Multiple Correspondence Analysis (6 p.)

A survey was administered to 201 first year students of engineering during the academic year of 2013-2014. To goal of this survey was to understand what are the factors influencing the success of a student in his first year at university. Variables examined are presented below:

- Attendance to class: *low, average, high*
- Number of study hours for exams: *less than 4h., between 4h. and 8h., more than 8h.*
- Active participation in class: *low, high*
- Participation to additional guidance: *low, high*

Use the picture and the eigenvalues (next page) to justify your answers.

(a) What is the total variance of the variables?

(b) How much of the total variance do the first three MCA components explain?

(c) Based on the picture, does it seem that students with average attendance to class are active in class? Justify!

(d) Based on the picture, does it seem that students with average attendance study more than 8 hours to the exam? Justify!

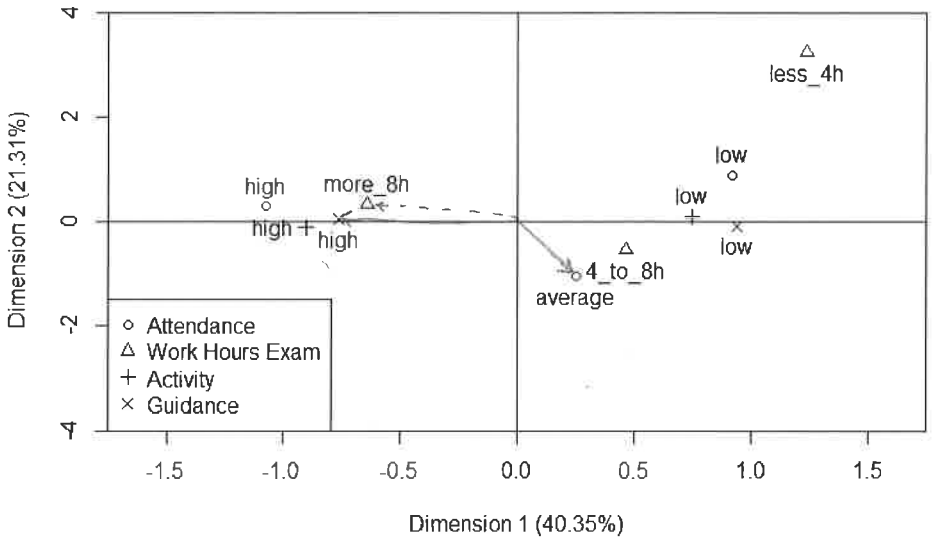

Table 1: Eigenvalues (rounded) associated with the MCA transformation:

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
0.61	0.32	0.21	0.17	0.11	0.08

3. Principal Component Analysis (6 p.)

   Let $x \in \mathbb{R}^{p\times 1}$ be a p-variate vector with mean μ and covariance matrix Σ. Let

   $$\Sigma = \Gamma\Lambda\Gamma^T,$$

   where the column vectors of Γ are the orthogonal eigenvectors of Σ, and Λ is a diagonal matrix having its diagonal elements in decreasing order. Let $y = \Gamma^T(x-\mu)$, and let $z = a^T x$, where $a \in \mathbb{R}^{p\times 1}$ and $a^T a = 1$. Show that the variance of the first element of the vector y is larger than or equal to the variance of z.

4. Depth functions (6 p.)

   According to Zuo and Serfling, depth functions should fulfill four general properties. State the four properties and explain (using 2-3 sentences) what they mean.

BONUS QUESTION (2 p.):

Let $\bar{x}$ denote the sample mean vector, and S denote the covariance matrix of the p-variate data

$$x_1, x_2, ..., x_n.$$

Let $A \in \mathbb{R}^{p \times p}$ be nonsingular, and let $b \in \mathbb{R}^p$. Consider now the data

$$Ax_1 + b, Ax_2 + b, ..., Ax_n + b.$$

Show that the sample mean vector of the revised data is $A\bar{x} + b$ and that the covariance matrix of the revised data is ASA^T.

Fun for all the math geeks:
http://spikedmath.com/
Check also the ARCHIVE.